Classification: Biological Sciences, Evolution, Biophysics and Computational Biology

Title: The Reconstruction and Evolutionary History of Eutherian Chromosomes

Short title: *Mammalian chromosome evolution*

Jaebum Kim[a1], Marta Farré[b1], Loretta Auvil[c], Boris Capitanu[c], Denis M. Larkin[b2], Jian Ma[d2], Harris A. Lewin[e2]

Author affiliation:

[a]Department of Biomedical Science and Engineering, Konkuk University, Seoul 05029, Korea

[b]Royal Veterinary College, University of London, London, NW1 0TU, UK

[c]Illinois Informatics Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[d]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[e]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA

[1]J. K. and M. F. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: dlarkin@rvc.ac.uk, lewin@ucdavis.edu, or jianma@cs.cmu.edu

Corresponding authors:

Harris A. Lewin, Department of Evolution and Ecology, University of California, Davis, CA 95616, USA, lewin@ucdavis.edu, Tel: +1 530-754-5098

Denis M. Larkin, Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London, London, NW1 0TU, UK, dlarkin@rvc.ac.uk, Tel: +44 (0)20 74685506

Jian Ma, Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA jianma@cs.cmu.edu, Tel: +1 412-268-2776

Keywords: chromosome evolution, ancestral genome reconstruction, genome rearrangements

Abstract

Whole genome assemblies of 19 placental mammals and two outgroup species were used to reconstruct the order and orientation of synteny blocks in chromosomes of the eutherian ancestor and six other descendent ancestors leading to human. For ancestral chromosome reconstructions, we developed a new algorithm (DESCHRAMBLER) that probabilistically determines the adjacencies of syntenic blocks using chromosome-scale and fragmented genome assemblies. The reconstructed chromosomes of the eutherian, boreoeutherian and euarchontoglires ancestor each included >80 percent of the entire length of the human genome, while reconstructed chromosomes of the most recent common ancestor of simians, catarrhini, great apes, and humans and chimpanzees included >90% of human genome sequence. These high coverage reconstructions permitted reliable identification of chromosomal rearrangements over ~105 million years (My) of eutherian evolution. Orangutan was found to have eight chromosomes that were completely conserved in homologous sequence order and orientation with the eutherian ancestor, the largest number for any species. Ruminant artiodactyls had the highest frequency of intrachromosomal rearrangements, while interchromosomal rearrangements dominated in murid rodents. A total of 162 chromosomal breakpoints in evolution of the eutherian ancestral genome to the human genome were identified; however, the rate of rearrangements was significantly lower (.80/My) during the first ~60 million years of eutherian evolution, then increased to greater than 2.0/My along the five primate lineages studied. Our results significantly expand knowledge of eutherian genome evolution and will facilitate greater understanding of the role of chromosome rearrangements in adaptation, speciation, and the etiology of inherited and spontaneously occurring diseases.

Significance Statement

Determining the order and orientation of conserved chromosome segments in the genomes of extant mammals is important for understanding speciation events, and the lineage-specific adaptations that have occurred during ~200 million years of mammalian evolution. In this paper, we describe the computational reconstruction of chromosome organization for seven ancestral genomes leading to human, including the ancestor of all placental mammals. The evolutionary history of chromosome rearrangements that occurred from the time of the eutherian ancestor until the human lineage was revealed in detail. Our results provide an evolutionary basis for comparison of genome organization of all eutherians, and for revealing the genomic origins of lineage-specific adaptations.

\body

Introduction

Chromosome rearrangements are a hallmark of genome evolution and essential for understanding the mechanisms of speciation and adaptation (1). Determining chromosome rearrangements over evolutionary time scales has been a difficult problem due primarily to the lack of high quality, chromosome-scale genome assemblies that are necessary for reliable reconstruction of ancestral genomes. For closely related species with good map-based assemblies, such as human, chimpanzee and rhesus, it is possible to infer most inversions, translocations, fusions and fissions that occurred during evolution by simple observational comparisons (2). However, for genome-wide comparisons that require resolving large numbers of rearrangements of varying scale, determining ancestral chromosomal states is challenging both methodologically and computationally due to the complexity of genomic events that have led to extant genome organizations, including duplications, deletions and reuse of evolutionary breakpoint regions (EBRs) flanking homologous synteny blocks (HSBs) (3, 4).

A variety of methods have been used for resolving the evolutionary histories of mammalian chromosomes, with limited success and resolution. For example, chromosome painting by fluorescent in situ hybridization (FISH) (5-8) was used to predict ancestral karyotypes dating back ~105 million years (My) to the ancestor of all eutherian (placental) mammals (9). While yielding an outline of the basic reconstructed karyotypes, FISH-based methods do not have sufficient resolution to permit accurate identification of EBRs, HSBs, and fine scale rearrangements. Low resolution methods also severely limit study of the relationship between chromosome rearrangements and structural variants, which are associated with adaptive evolution and the presence of EBRs (4, 10, 11). Thus, a distinct advantage of resolving EBRs at high resolution is that sequence features within them can be interrogated for genes that may be associated with lineage-specific phenotypes. This is an important motivation for creating finer scale ancestral chromosome reconstructions (10, 12, 13).

Several algorithms have been developed to reconstruct the order and orientation of synteny blocks in common ancestors by using DNA sequence-level syntenic relationships among genomes of extant species (14). These methods use syntenic fragments (SFs) constructed from whole-genome sequence alignments as input to infer the order and orientation of the SFs in a specific target ancestor. Different algorithmic approaches are used by the different reconstruction algorithms. For example, the MGR algorithm uses a heuristic approach to solve the problem of sorting by reversals (inversions) for multiple genomes based on rearrangement distance. inferCARs finds the most parsimonious scenario for the history of SF adjacencies and then greedily connects the adjacencies into contiguous ancestral regions (CARs). The MGRA algorithm utilizes multiple breakpoint graphs based on SFs in descendent species to infer the ancestral order of SFs, while ANGES uses "consecutive one property" to cluster and order SFs in a target ancestor. However, these methods have been used to reconstruct just a small number of ancestral mammalian genomes, primarily because there are a very limited number of chromosome-scale whole genome assemblies (4, 12). Also, it has not been shown whether these existing algorithms for reconstructing chromosome organization are suitable for fragmented assemblies produced by next-generation sequencing technologies.

Examples of mammalian genome reconstructions reveal the limitations of earlier datasets. Murphy et al. (15) applied MGR to human, cat, cow and mouse genome maps and assemblies to reconstruct the chromosome organization of the boreoeutherian ancestor, which lived ~97.5 million years ago (Ma) (9). Subsequently, the boreoeutherian, ferungulate, carnivore and other ancestral genomes were reconstructed using MGR, combining physical maps and sequence information from eight species representing five mammalian orders (4). Twenty-three pairs of autosomes plus sex chromosomes were predicted for the boreoeutherian ancestor, but sequence coverage as measured against the human genome was only about 50% (4), resulting in limited definition and accuracy of both large scale and fine scale (<1.0 Mbp) chromosome rearrangements. In a later study (3), inferCARs was used to reconstruct CARs of the boreoeutherian ancestor that were generally consistent with chromosome painting results, but the reconstruction was limited and coarse due to the small number of descendent species used. In addition, there were studies using genes as markers to reconstruct the order and orientation

of HSBs in the boreoeutherian ancestor (e.g., (16)), but it is unclear how much gene-based reconstruction represents the ancestral reconstruction using whole-genome sequencing data. Therefore, although these recent results were an improvement over earlier work, missing information from other mammalian orders and use of low resolution maps contributed to the reduced coverage, thus limiting the potential usefulness of the reconstructions for evolutionary and functional analysis.

Despite some recent improvements in reconstruction algorithms (3, 14, 17, 18), the field has been more or less stagnant for the past decade because of the paucity of new genome assemblies suitable for ancestral reconstructions. In this paper, we introduce a method, called DESCHRAMBLER, which uses SFs constructed from whole-genome comparisons of both high quality chromosome-scale and fragmented assemblies. The method is an extension of the algorithm for Reference Assisted Chromosome Assembly, or RACA (19), which implements a probabilistic framework to predict adjacencies of SFs in a target species. DESCHRAMBLER has the flexibility to handle chromosome-level and scaffold assemblies, and is scalable to accommodate a large number of descendent species. In the present study, we applied DESCHRAMBLER to sequenced genomes of 21 species that included representatives of 10 eutherian orders. Results reveal a detailed picture of chromosome rearrangements that occurred during ~105 My of eutherian evolution.

Results

Chromosome reconstruction for seven eutherian ancestors of H. sapiens

The chromosome organizations of seven common ancestors in the lineage leading to human were reconstructed using genome assemblies of 19 extant eutherian species and two outgroup species, one a marsupial and one a bird (Supplementary Table S1). Genomes were selected on the basis of their availability in public databases, quality of genome assembly and taxonomic order (Fig. 1, Supplementary Table S1; see Materials and methods). The set of species contains representatives of ten orders of eutherian mammals: primates (human, chimpanzee, orangutan,

rhesus, and marmoset), rodentia (mouse, rat, and guinea pig), lagomorpha (pika), cetartiodactyla (cattle, goat, and pig), perissodactyla (white rhinoceros and horse), carnivora (dog), eulipotylpha (shrew), proboscidea (elephant), sirenia (manatee), and afrosoricida (tenrec), and two outgroup species to eutheria (opossum and chicken). Among the 21 genome assemblies, 14 were chromosome-level, and the remaining seven were assembled as sequence scaffolds with N50 ranging 14.4-46.4 Mbp. The number of scaffolds in fragmented assemblies ranged from 2,352 (elephant) to 12,845 (shrew). Total sequenced genome size varied from 1 Gbp (chicken) to 3.5 Gbp (opossum; Supplementary Table S1). For reconstruction of ancestral chromosomes, the human genome was used as the reference for alignments because of the relative quality of the assembly, and because we focused reconstructions on the evolution of lineages leading to human. Two resolutions (500 and 300 Kbp minimum breakpoint distance in the human genome) were selected to create the SFs that are used by the DESCHRAMBLER reconstruction algorithm as input. Herein, we made our interpretations on the basis of 300 Kbp resolution; results at 500 Kbp (Supplementary Tables S5-S7) were used for comparison to help resolve discrepancies with FISH data and to better understand differences in breakpoint rates along the different lineages.

The number of ancestral predicted chromosome fragments (APCFs) ranged from 30 in the common ancestor of great apes, to 35 in the common ancestor of human and chimpanzee (Table 1). The SFs of each ancestor were defined using only the descendant species from the corresponding ancestral node (with the rest as outgroup species). Therefore, SFs of the more ancient ancestors contained homologous genomic regions from a larger number of descendant species than the more recent ancestors. This accounts for the greater number of smaller SFs and the smaller total size of APCFs in more ancient ancestors. However, the difference in APCF sizes among ancestors was minimized by allowing missing coverage in SF definitions for a small number of descendent genomes (Materials and methods). The APCFs of the simian, catarrhini, great apes, and common ancestor of human and chimpanzee cover more than 90% of the human genome, whereas the eutherian, boreoeutherian, and euarchontoglires APCFs each cover more than 80% of the human genome.

Comparison with existing ancestral genome reconstruction algorithms

Three existing tools for ancestral chromosome reconstruction, ANGES (18), inferCARs (3), and MGRA (17), were used to compare results obtained with DESCHRAMBLER. For a fair comparison, the same sets of SFs were used as input to the above three tools, and the predicted adjacencies of SFs in the seven target ancestors were compared. The number of APCFs obtained with DESCHRAMBLER ranged from 30 in the common ancestor of great apes, to 35 in the common ancestor of human and chimpanzee (Supplementary Table S2). The other three tools produced larger numbers of APCFs for the eutherian ancestor, which are due apparently to the increased number of descendent species with scaffold assemblies having unclear definition of chromosome ends (Supplementary Table S1). Other than for the eutherian ancestor, ANGES consistently produced the fewest APCFs, whereas MGRA produced very large numbers of APCFs, particularly for the most distant common ancestors to human. Comparison of predicted SF adjacencies among the four tools showed that the results obtained with DESCHRAMBLER were highly similar to those of ANGES and inferCARs (Jaccard similarity coefficient > 0.8; Supplementary Fig. S3). Results from DESCHRAMBLER and inferCARs were the most similar for all of the seven reconstructed ancestral genomes, whilst the greatest discrepancies were found between MGRA and the other tools (Supplementary Fig. S3).

Comparison with FISH-based reconstructions of ancestor chromosomes

We compared the eutherian, boreoeutherian and simian ancestral karyotypes determined by FISH (6, 8, 20) with those obtained using DESCHRAMBLER and three additional tools (see Materials and methods for details). In this evaluation, we focused on interchromosomal rearrangements using human chromosomes as a reference. For example, there are seven fusions of human chromosomes found in the eutherian and boreoeutherian ancestors, and two fusions of human chromosomes in the simian ancestor (Table 2). DESCHRAMBLER agreed with FISH data in 12/16 cases, thus outperforming the other three tools. In 3/4 cases where FISH data and DESCHRAMBLER disagreed, DESCHRAMBLER partially predicted the interchromosomal rearrangements. For example, in the reconstructed chromosomes of the eutherian ancestor, the descendent homologs HSA8p and parts of HSA4 were predicted to be fused by DESCHRAMBLER, but joining of HSA8p to another segment of what is now HSA4q was not detected (Supplementary

Table S7). Similarly, in the reconstructed chromosomes of the eutherian and boreoeutherian ancestors, the descendent homologs HSA12pq and HSA22q were predicted to be fused by DESCHRAMBLER, but joining to what is now HSA10p was not detected. However, in the eutherian and boreoeutherian ancestral genomes, the fusion of HSA10p to 12pq-22q is weakly supported in FISH-based reconstructions (6). ANGES was the next best performer with 11 agreed cases. MGRA produced the lowest agreement with the FISH-based reconstructions because of the highly fragmented nature of its APCFs in the three ancestors used in this evaluation (Supplementary Table S2).

One large eutherian APCF produced by DESCHRAMBLER was not supported by FISH data. This APCF (see EUT1, Supplementary Table S3) joined what is now all of HSA4 and HSA13, and parts of HSA8 and HSA2. The organization of this large APCF partially agrees with the ancestral eutherian chromosome formed by what is now HSA8p and HSA4pq as predicted by chromosome painting (Table 2) (4). It is noteworthy that both eutherian ancestral adjacencies involving homologs of HSA8 and HSA2, and HSA2 and HSA13, have a high DESCHRAMBLER score (> 0.999) and are spanned by one chromosome or scaffold in the Afrotherian and outgroup species. In addition, ANGES predicted the same ancestral configurations in the eutherian ancestor, while inferCARS split it into two APCFs (Supplementary Table S7). Therefore, there are multiple lines of evidence to support the EUT1 adjacencies in the eutherian ancestral genome, although there are discrepancies among the reconstruction methods and at different resolutions (Supplementary Table S7). Finally, the fusion of two ancestral chromosomes homologous to HSA7 was predicted by BAC-FISH to occur in the ancestral catarrhini genome (20), while DESCHRAMBLER placed it in the simian ancestor. High-confidence FISH-based chromosomal configurations in each ancestor were incorporated into the final reconstruction of ancestral genomes predicted by DESCHRAMBLER (Supplementary Table S3).

Evolutionary breakpoints and chromosome rearrangements

At 300 Kbp resolution, we detected 162 chromosomal breakpoints that occurred during 105 My of mammalian evolution, from the eutherian ancestor's genome to the human genome (Fig. 1; Supplementary Table S4). Six breakpoints occurred on the branch from eutheria to

boreoeutheria, which correspond to three fissions, one inversion, and one complex rearrangement. There were nine breakpoints in the euarchontoglires ancestor's genome in comparison to the boreoeutherian ancestor's genome, resulting in one fusion, two fissions, three inversions, and two complex rearrangements. The number of rearrangements increased during evolution from the euarchontoglires ancestor to the more recent ancestors. Among them, the largest number of rearrangements (N=38) occurred from the euarchontoglires ancestor to the simian ancestor, producing 47 evolutionary breakpoints. Mostly inversions and complex rearrangements were observed during the evolution of the eutherian ancestor to human, whereas fusions and fissions were less prevalent.

We next examined the number of chromosome breakpoints in terms of divergence time from common ancestors (Supplementary Tables S4-S6). At 300 Kbp resolution the lowest breakage rate was 0.80/My, occurring from the eutherian ancestor to the boreoeutherian ancestor (FDR P <0.05). The breakage rate was lower on the branch from the euarchontoglires ancestor to the simian ancestor (0.98/My, FDR P <0.05), and higher on the branch from the common ancestor of great apes to the common ancestor of human and chimpanzee (3.59/My, FDR P <0.10). During the evolution of primate ancestors to extant primate genomes, breakage rates in the lineages leading to rhesus and chimpanzee were significantly higher than along other branches (4.19/My, FDR P <0.05, and 6.21/My, FDR P <0.05 respectively) and was lower in the lineage leading to orangutan (1.08/My, FDR P <0.05). We then compared the results obtained at 300 Kbp resolution with those obtained at 500 Kbp resolution (Supplementary Tables S5-S6). Although breakage rates were consistently lower at 500 Kbp resolution, levels of statistical significance were consistent for all comparisons except for orangutan.

We then investigated possible causes of the differences in chromosome breakage rates at 300 Kbp and 500 Kbp resolution. The number of SFs below the 500 Kbp and 300 Kbp thresholds were compared by counting the number of SFs at 300 Kbp resolution corresponding to each branch and then correlating these results with the amount of breakpoint increase (Supplementary Fig. S4). There was a high linear correlation between the two measures in terms of both the absolute number and the fraction of small SFs. Thus, the increase in breakpoints was

mostly attributed to smaller scale rearrangements between 300 Kbp and 500 Kbp because inversions and complex rearrangements were observed in higher numbers at 300 Kbp resolution (Supplementary Tables S4-S5).

Evolutionary history of the eutherian ancestor's genome

A complete summary of the evolutionary history of each reconstructed ancestral eutherian chromosome is presented in Fig. 2, the Supplementary Text and Fig. S2. An integrated summary of results with emphasis on chromosome rearrangements in the lineage leading to human is presented below.

Comparative analysis of reconstructed chromosomes of the eutherian ancestor revealed that a majority were highly stable in both the boreoeutherian and euarchontoglires ancestral genomes (Fig. 2; Supplementary Fig. S2). The exceptions to this pattern were the descendent homologs of EUT1 and EUT6, which were separated by fission into three and two chromosomes, respectively, in the boreoeutherian ancestor's genome. Another exception was the descendent homolog of EUT13, which gained a ~10 Mbp inversion in the boreoeutherian ancestor. The descendent homolog of EUT18 gained large inversions in the euarchontoglires ancestor's genome but was maintained as a single chromosome (Supplementary Fig. S2).

In the reconstructed simian ancestor's genome, 15/21 eutherian ancestor chromosomes were conserved as a single chromosome, of which five underwent intrachromosomal rearrangements (Fig. 2). Among the 15 conserved full-chromosome syntenies, 13 were conserved as single chromosomes or chromosome blocks within larger chromosomes in human, chimpanzee and orangutan, the largest number for any extant species. Two descendent homologs of eutherian ancestor chromosomes with synteny conserved in the simian ancestor's genome underwent interchromosomal rearrangements later in the primate lineage; EUT2 (a fission in the catarrhini ancestor) and EUT7 (a fission in the ancestor of great apes) (Supplementary Fig. S2). In comparison, 12 eutherian ancestor chromosomes have homologs in pig with completely conserved synteny, the greatest number for any extant non-primate species in our analysis;

however, 11 of these underwent intrachromosomal rearrangements. The species with the fewest conserved chromosomes relative to the eutherian ancestor was mouse, with three.

No additional rearrangements in evolutionary stable eutherian ancestor chromosomes (i.e., those without internal rearrangements) were introduced in the reconstructed catarrhini ancestor genome as compared with the simian ancestor. However, three descendent homologous chromosomes of the eutherian ancestor (EUT8, EUT9, and EUT17) underwent lineage-specific complex rearrangements in the human-chimpanzee ancestor (Fig. 2; Supplementary Fig. S2). We found six eutherian ancestral chromosomes (EUT4, EUT5, EUT12, EUT14, EUT20 and EUTX) that had no interchromosomal or intrachromosomal rearrangements during ~98.4 million years of evolution until the common ancestor of human and chimpanzee (Fig. 2; Supplementary Fig. S2). Among all extant species studied, orangutan was found to have the largest number of chromosomes (N=8) that were completely conserved in SF order and orientation compared with homologs in the eutherian ancestor. In the human lineage, the descendent homolog of EUT14 underwent a large (~12 Mbp) inversion (Fig. 3), while in chimpanzee its structure follows the ancestral eutherian configuration.

The largest number of intrachromosomal rearrangements in the primate lineage occurred in the evolution of EUT15 (Fig. 3), with the majority of these events dating to the simian ancestor, and additional rearrangements occurring later in the catarrhini and in the human-chimpanzee ancestor's genomes. Both the human and chimpanzee genomes exhibit additional rearrangements in the descendent homologs of EUT15 (HSA17 and PTR17, respectively). In contrast, EUT15 was found completely conserved in the mouse and horse genomes (Fig. 3), whereas the cattle and goat genomes contained just one large inversion in their descendent homologs of EUT15.

Although EUTX was highly conserved among primates, artiodactyl species had significant numbers of X chromosome inversions, whereas the order and orientation of EUTX SFs in horse (a perissodactyl) were conserved. There are small inversions and/or interchromosomal rearrangements observed in the X chromosomes of murid rodents, dog (a carnivore), cattle, and

other lineages, but assembly errors cannot be ruled out as causing at least some of these apparent rearrangements.

Overall, 537.5 Mbp of the reconstructed eutherian ancestor's genome (21.8% of total eutherian genome size) lack both interchromosomal and intrachromosomal rearrangements, and an additional 798.5 Mbp (32.4% of total genome size) of the eutherian ancestor chromosomes had intrachromosomal but no detectable interchromosomal events during evolution to the human genome (Supplementary Table S9). The remaining 45.8% was found in reconstructed eutherian chromosomes that underwent interchromosomal and interchromosomal rearrangements. This compares to 3.8% and 2.6% maximum eutherian ancestor genome coverage observed for chromosomes with no interchromosomal or intrachromosomal rearrangements, and 36.5% and 7.0% maximum coverage for intrachromosomal-only rearrangements in artiodactyl and murid genomes, respectively (Supplementary Table S9). Thus, compared with the reconstructed eutherian genome, the primate lineage tends to have a larger fraction of genomes in unrearranged synteny blocks as compared to other eutherian lineages.

Unassigned APCFs

DESCHRAMBLER produced two small chromosomal fragments, Un29 (1Mbp) and Un30 (0.5 Mbp) that were not joined to any reconstructed chromosomes in the eutherian ancestor genome (Supplementary Fig. S2). These fragments must have been produced by multiple independent rearrangements (i.e., reuse breakpoints, (11)) in several mammalian clades. It is likely that in the lineage leading to primates these fragments were adjacent and located at the telomeric region of the EUT1 homolog. In the simian and later in the catarrhini ancestral genomes, several inversions separated Un29 and Un30, which are found about 10 Mbp apart on HSA1. Thus, independent chromosomal rearrangements apparently reorganized these fragments in artiodactyl, rodent and perissodactyl lineages, indicating that these APCFs are bounded by highly dynamic intervals in eutherian chromosomes.

Discussion

Chromosomes of seven ancestral genomes along the 98.4 million-year lineage from the ancestor of all placental mammals to the common ancestor of humans and chimpanzees were reconstructed using the DESCHRAMBLER algorithm. Seven of the extant species had sub-chromosomal, scaffold-level assemblies that were effectively used by DESCHRAMBLER to reconstruct ancestral chromosome fragments and to identify lineage-specific chromosome breakpoints. The reconstructions were made using genomes of extant species from 10/19 orders of eutherian mammals representing the Laurasiatheria, Afrotheria, and Euarchontoglires superorders. Although Xenarthra was not represented, species from these three superorders permitted reconstruction of the eutherian, boreoeutherian, and euarchontoglires ancestor's chromosomes at high resolution as compared with the earlier FISH-based reconstructions (6, 8, 20). The ancestral reconstructions far surpassed the quality of previous map and sequence-based reconstructions in terms of the number of descendent species included, coverage of ancestor genomes relative to the human genome, and the number of ancestors in the evolutionary path to the human genome (3, 4), thus providing novel insights into eutherian and primate genome evolution.

The choice of a reference genome is critical for the completeness of chromosome reconstructions because the reference is used as a backbone to find orthologous chromosomal regions in different species using whole-genome sequence alignment, and to construct SFs that are shared between species. It is noteworthy that our reconstruction algorithm itself does not bias toward any descendant genome, but the reference genome has an impact on the SFs that we use for the reconstruction. The human genome was used as a reference because it considered to have the highest quality assembly among the mammals, and because all ancestors targeted for genome reconstruction were ancestral to human. In addition, assembly quality is also important for overall accuracy and completeness of the SFs. To reduce the complications in reconstruction introduced by extensively fragmented genome assemblies, we selected species with assemblies that have N50 scaffold size >14 Mbp and that could be aligned against more than 80% of the

reference human genome. Because we only used one reference genome in the present work for defining SFs, it is possible that some ancestral sequences that are not present in the human genome were omitted in the reconstructions. It would be useful to develop SF construction methods that consider multiple reference genomes, similar to what has been done for bacterial genomes (21). In addition, recent developments in long read sequencing technologies (22), genome scaffolding (23-25), and comparative and integrative mapping (19, 26) produce higher quality assemblies that approach whole chromosomes. These methods are now cost effective relative to creating high-density BAC maps, linkage maps and radiation hybrid maps (12), and will be useful for providing higher quality SFs that may greatly facilitate the understanding of chromosome evolution using ancestral genome reconstruction methods.

For ancestral genome reconstruction, DESCHRAMBLER takes into account clade-specific or species-specific insertions and deletions. If the SFs are constructed by requiring orthologous chromosomal regions from all descendent species, the genome of their common ancestor would not be well covered, especially when the genomes of the descendent species are highly diverged or the assemblies are incomplete. To address this issue, SFs were created without the above constraint of the inclusion of all orthologous genomic regions. Instead, all possible SFs were first created with a different number of genomic regions of descendent species, and then candidate SFs for each target ancestor were chosen by a parsimony algorithm based on the presence and absence of orthologous genomic regions in each descendent species. To take advantage of these new SFs, the reconstruction algorithm must be able to utilize them. Most existing algorithms, such as ANGES, inferCARs, and MGRA, were developed using the assumption of strict constraint on orthologous regions in SFs that orthologous regions from all descendent species must exist in an SF. However, DESCRHRAMBLER is more flexible in utilizing SFs when some of the species have deletions of genomic regions or there is missing data. This is one of the reasons why DESCHRAMBLER outperformed other existing tools in the reconstruction of the oldest (EUT) ancestor.

After incorporating high-confidence FISH-based chromosomal configurations in each ancestor, we deduced an ancestral eutherian karyotype having 2n=44 chromosomes (assuming a separate

Y chromosome). This number is lower than FISH-based inferences of 2n=46 (5, 6, 8, 27, 28), and is due to the reconstructed EUT1 (ascendant homolog of HSA13, HSA2, HSA4 and HSA8) and EUT6 (partially homologous to HSA7 and HSA10). Our results are in agreement with previous studies that used FISH-based and sequenced-based methods to deduce the ancestral boreoeutherian karyotype to have 2n=46 chromosomes (3, 5, 6, 27, 28). We also deduced an ancestral catarrhini karyotype of 2n=46, an ancestral great apes karyotype of 2n=48, and 2n=48 for the human-chimpanzee ancestor, which all agree with results from chromosome painting and BAC-FISH experiments (20, 28).

In the simian ancestor (the ancestor of Old World and New World monkeys), we reconstructed an ancestral karyotype with 2n=46 chromosomes. This number is lower than obtained with FISH-based methods, which inferred 2n=48 (5, 28) or 2n=50 (20). The main differences are SIM7 (homolog to HSA7) and SIM10 (homolog to HSA10), where DESCHRAMBLER created one ancestral chromosome for each chromosome, while FISH data consistently supported reconstruction of HSA7 and HSA10 each into two fragments (5, 20, 28). In summary, the diploid numbers of ancestor genomes deduced by DESCHRAMBLER were very similar to the results of previous reconstructions. Additional high quality genome assemblies will help to resolve remaining discrepancies.

We have clearly demonstrated that each eutherian chromosome has a unique evolutionary history in the different mammalian lineages, and that many ancestral eutherian chromosomes were stable in descendent lineages, with relatively few large-scale rearrangements in the ancestral genomes leading to human. Among the primate species included in the analysis, more than 100 putative breakpoints were detected during evolution from the simian ancestor to marmoset, and from the catarrhini ancestor to rhesus (Supplementary Table S6), thus indicating an accelerated rate of evolution in these non-human primates during the past 43 million years (see below). Although the time from the great ape ancestor to the common ancestor of human and chimpanzee has a relatively short branch length (9.2 My), there were 14 inversions and 10 complex rearrangements (i.e., a combination of inversions and putative transpositions) assigned to that branch, which also gives the highest breakpoint rate on that particular lineage. For

comparison, we looked at the breakpoint rates from these ancestral nodes along the lineages to other non-human descendant species. We found that the branch from the great ape ancestor to orangutan has the lowest breakpoint rate (1.08/My) as compared to other branches (Supplementary Table S6) and the result was consistent when we used 500 Kbp as the SF resolution. This suggests an overall higher chromosomal rearrangement rate on the branch from the great ape ancestor to the ancestor of human and chimpanzee, but a much slower rate from the great ape ancestor to orangutan. In addition, our results refined the previously reported comparison between the orangutan genome and human-chimpanzee ancestor (29), where 40 rearrangements events were identified at 100 Kbp resolution. Regardless of varying rates of rearrangements within different primate lineages, comparison with other mammalian orders included in this work indicates that the primate ancestor and several descendent species' genomes contain the largest fraction of descendent homologs of eutherian ancestor chromosomes either totally conserved or affected by intrachromosomal rearrangements only. This suggests that the small insectivorous and scansorial common ancestor of all existing placental mammals (30) had chromosome structures highly resembling those of some contemporary primates (e.g., orangutan and human).

The breakpoint rate in the lineage leading to chimpanzee was almost three-fold higher than in the lineage leading to human at 300 Kbp resolution (6.21/My and 1.97/My, respectively), and more than four-fold greater at 500 Kbp resolution (Supplementary Tables S4-S6). These results indicate true differences in the rate of chromosome evolution in the lineages leading to humans and chimpanzees. Interestingly, the number and the rate of breakpoints in orangutan chromosomes remained constant for the two breakpoint resolutions, indicating few if any rearrangements that are in the 300-500 Kbp range in this species. On the basis of the above analyses we recommend that ≥300 Kbp resolution be used to analyze chromosomal rearrangements that affect the synteny and order of homologous sequences in order to avoid most false breakpoints introduced by assembly errors, as well as segmental duplications and copy number variants. However, the use of multiple breakpoint resolutions can be advantageous when the goal is to draw more accurate and comprehensive conclusions from many descendent species to reveal the interplay between large-scale rearrangements and finer resolution genomic

changes (including duplications). Therefore, there should be additional efforts to enhance reconstruction algorithms to effectively aggregate results at different resolutions of breakpoint intervals.

The analysis of chromosome evolutionary breakpoint rates yielded results that are generally consistent with Murphy et al. (2) who found slow rates of chromosome evolution in mammals prior to the K-P boundary, which corresponds to the massive extinction event that led to the disappearance of the dinosaurs (except for birds) and the eventual rise of mammals. We also found an accelerated rate of chromosome rearrangements in primate ancestors, specifically along the branch leading to the common ancestor of humans and chimpanzees. The significance of these findings is unclear, but might be related to differences in genomic architecture, repetitive elements, and changes in the environment that are known to cause chromosome rearrangements (11). Assembly errors may also cause an increase in the apparent rate of rearrangements, and these must be excluded prior to drawing conclusions. One way to approach this problem is to compare breakpoint rates at different resolutions. Fewer breakpoints are expected at lower resolution, but the relative differences in rates should be stable. Consistent with this expectation, we found a linear correlation between the number of SFs <500 but >300 Kbp and the number of breakpoint differences at 300 Kbp and 500 Kbp (Supplementary Figure S4). From additional analysis, we also observed that the small SFs contributed to creating rearrangements involving inversions and other complex rearrangements (Supplementary Tables S4-S5). Breakpoints generated by these smaller SFs are either the footprint of *bona fide* structural rearrangements, or they may be artifacts produced by misassembled sequences. For example, previous studies revealed problems in the rheMac2 assembly version of the rhesus genome (31-33), which is one of the species showing a large discrepancy of the number and the rate of breakpoints at the two resolutions. Even though we used a more recent version of the rhesus genome (rheMac3), it is not clear whether all of the assembly problems in the previous version were completely fixed.

The reconstructed events of chromosome evolution in multiple ancestral genomes leading to human permitted assignment of breakpoints to different branches in the phylogeny. Such

information can be useful for further analysis of the potential functional roles of chromosomal rearrangements in eutherian evolution. Earlier work reported an association between evolutionary breakpoints and gene functions that may contribute to lineage- and species-specific phenotypes (11, 34). More recently, such association analysis has been extended to understanding the relationship between chromosome rearrangements and non-coding function elements of the genome such as open chromatin regions (16). In the present study, we found two small APCFs of the eutherian ancestor (Un29 and Un30) that were not assigned to specific ancestral chromosomes due to the fact that these two fragments were flanked by breakpoint regions with independent reuse in different eutherian lineages. If we examine the gene content within these EBRs using the human genome as a reference, we find them to contain multiple paralogs of zinc finger and olfactory receptor genes, which have been found previously to be enriched within EBRs (11, 35), are associated with adaptive evolution (36, 37), and may promote rearrangements by non-allelic homologous recombination (e.g., (38)). Specifically, the fragment Un29 is flanked by zinc finger genes ZNF678 and pseudogene ZNF847P at one end, and three histone genes (HIST3H3, HIST3H2BB, HIST3H2H) at the other (data not shown, but can be visualized on the UCSC Genome Browser). Among the other 17 genes found within Un29 are several gene family members, including *WNT3A*, *WNT9A*. It has been shown that small changes in expression of WNT genes can result in a radical alteration of body plan (39). In the human genome, Un30 is flanked by three zinc finger genes (ZNF670, ZNF669, ZNF124), one additional zinc finger gene (ZNF496), and three olfactory receptor genes (OR2B11, OR2W5, OR2C3). Because chromosome rearrangements are known to affect regulation of gene expression (40), these data suggest that reuse of evolutionary breakpoint sites near this fragment in multiple clades could be a contributor to producing new variation in gene content and gene expression. With additional mammalian genomes being sequenced, our genome reconstruction approach has the potential to provide the foundation for a more comprehensive evolutionary analysis to improve understanding of the relationship between genome rearrangements, functional elements (both coding and non-coding), and adaptive traits.

Reconstruction of the chromosomes of seven descendent genomes, from the eutherian ancestor to human, is an excellent example of what can be achieved by applying similar analysis to other

clades. The recent advances in long read technology and scaffolding techniques will enable more rapid production of assemblies that are suitable for accurate identification of lineage-specific breakpoints, which are the basis for high quality ancestral chromosome reconstructions. Thus, in the near future, it will be possible to reconstruct genomes at the key nodes of all mammalian lineages, and to explore the nature of chromosome rearrangements that occurred during more recent radiations. As previously shown, karyotypes, physical maps and whole genome sequences with precise locations of centromeres and telomeres also add important information for understanding chromosome evolution, and for understanding the relationship between chromosome rearrangements, EBRs, cancers, and inherited human diseases (2, 41). Together with improved tools for aligning, comparing and visualizing large numbers of genomes, these new chromosome-scale assemblies will offer unparalleled opportunities to study the mechanisms and consequences of chromosome rearrangements that have occurred during mammalian evolution. With efforts such as those to sequence 10,000 vertebrate genomes (42), it will be possible to extend reconstructions deeper into evolutionary time, and thus provide a more detailed picture of chromosome evolution in other vertebrate classes. Ultimately, it should prove possible to determine the ancestral eukaryote chromosome organization, and to create a new chromosome nomenclature system that is based on evolutionary principles.

Materials and methods

Data

The pairwise genome sequence alignments (chains and nets) among 21 genome assemblies using the human genome as reference were downloaded from the UCSC Genome Browser (43) or directly constructed by using an alignment pipeline based on lastz (44) with the chain/net utilities from the UCSC Genome Browser. The genomes used were: human (*Homo sapiens*, GRCh37/hg19), chimpanzee (*Pan troglodytes*, CSAC 2.1.4/panTro4), orangutan (*Pongo pygmaeus abelii*, WUGSC 2.0.2/ponAbe2), rhesus (*Macaca mulatta*, BGI CR_1.0/rheMac3), marmoset (*Callithrix jacchus*, WUGSC 3.2/calJac3), mouse (*Mus musculus*, GRCm38/mm10), rat (*Rattus norvegicus*, RGSC 5.0/rn5), guinea pig (*Cavia porcellus*, Broad/cavPor3), pika (*Ochotona princeps*,

OchPri3.0/ochPri3), cattle (*Bos taurus*, Baylor Btau_4.6.1/bosTau7), goat (*Capra hircus*, CHIR_1.0/capHir1), pig (*Sus scrofa*, SGSC Sscrofa10.2/susScr3), white rhinoceros (*Ceratotherium simum*, CerSimSim1.0/cerSim1), horse (*Equus caballus*, Broad/equCab2), dog (*Canis lupus familiaris*, Broad CanFam3.1/canFam3), shrew (*Sorex araneus*, Broad/sorAra2), elephant (*Loxodonta africana*, Broad/loxAfr3), manatee (*Trichechus manatus latirostris*, Broad v1.0/triMan1), tenrec (*Echinops telfairi*, Broad/echTel2), opossum (*Monodelphis domestica*, Broad/monDom5), and chicken (*Gallus gallus*, ICGSC Gallus_gallus-4.0/galGal4). The tree topology of these 21 species was based on the tree used to align 45 vertebrate genomes with human in the UCSC Genome Browser, and branch lengths were estimated based on TimeTree (45).

Ancestral genome reconstruction algorithm

We developed a new method, called DESCHRAMBLER, to reconstruct the order and orientation of SFs in eutherian ancestral genomes. The workflow of the method is shown in Supplementary Fig. S1. The algorithm starts with the construction of syntenic fragments (SFs). Using a chromosome evolution model-based probabilistic framework, DESCHRAMBLER computes the probabilities of pairs of SFs being adjacent in a target ancestor based on the order and orientation of SFs in descendent as well as outgroup species. The SFs and their degree of adjacency in the target ancestor are next represented as a graph, which is used to estimate the most likely paths of SFs. The paths represent the order and orientation of SFs in the target ancestor. Details of each step are presented below.

Construction of syntenic fragments (SFs)

For each ingroup species, genomic blocks, which are matched to the nets of pairwise alignments with a reference, were mapped on reference genome sequences. The nets of length greater than a given threshold (resolution) were used, and colinear genomic blocks were merged together. After finishing this step for every ingroup species, the reference genome sequences together with the mapped genomic blocks of the other species were split at the boundaries where there were breaks in genomes of at least one species. Then aligned genomic blocks of outgroup species were added to each fragment, resulting in SFs. Not all SFs have genomic blocks from all ingroup

species, and therefore not all SFs were used in reconstruction. The SFs were used in reconstruction if the genomic blocks in the SF were predicted to share a common ancestral block in a target ancestor by using a parsimony algorithm that minimizes the number of state changes in intermediate ancestors to account for the presence and absence of blocks in extant species.

Computation of SF adjacency probabilities in a target ancestor

Given input SFs, their order and orientation in each ingroup and outgroup species are collected, which are used as the SF adjacency information in extant species. The probabilities of pairs of these SFs being adjacent in a target ancestor are computed from their adjacencies in extant species based on the probabilistic framework used in the RACA algorithm (19). The basic idea of the probabilistic framework is to calculate the posterior probability of pairs of SFs b_i and b_j being adjacent in the target ancestor by multiplying two posterior probabilities: b_i precedes b_j, and b_j succeeds b_i. The two posterior probabilities were calculated by using the Felsenstein's algorithm for likelihood (46) and the extended Jukes-Cantor model for breakpoints (47). More details can be found in (19).

Prediction of the order and orientation of SFs in a target ancestor

The probabilities of SF adjacencies in a target ancestor are used to construct a SF graph $G(V, E)$, which is an undirected graph with a set of vertices V representing SFs, and a set of edges E connecting vertices whenever there is an adjacency probability between two vertices. Each SF is expressed by using two vertices representing the head and tail of a SF. This is required because one SF can be connected to either the head or tail of another SF. Each edge has a weight representing the probability of adjacency between two connected vertices, and the head and tail vertices of the same SF always have the highest probability, 1.0. From the constructed SF graph, a greedy algorithm is used to predict the order and orientation of SFs in the target ancestor by incrementally merging two adjacent SFs according to the descending order of their edge weights, which is followed by the construction of lists of adjacent SFs. All SF adjacencies with a probability > 0 were used in the reconstruction for seven eutherian ancestors.

Refinement of predicted SF adjacencies

Weak SF adjacencies, which are (i) supported by just one ingroup species without any support from outgroup species or (ii) not supported by any ingroup species, are split. Then among the collection of lists of adjacent SFs, any two lists $L_1(a_1, \ldots, a_n)$ and $L_2(b_1, \ldots, b_m)$, where the adjacency between two SFs a_n and b_1 has a weight and is unambiguously supported by the parsimony algorithm by considering their adjacencies in descendent species, are merged to create a new list of adjacent SFs $L_{12}(a_1, \ldots, a_n, b_1, \ldots, b_m)$. This process repeats until no newer list of SFs is created. We note that L_1 and L_2 can be merged by four different ways (L_1 L_2, L_1 $-L_2$, $-L_1$ L_2, and $-L_1$ $-L_2$, where the '-' symbol represents a reversal of a list). Therefore, if there is more than one way to meet the above criteria, the one with the maximum adjacency weight is chosen.

Many of the APCFs initially reconstructed using DESCHRAMBLER (and with the other tools) are fragments of chromosomes. For example, the number of APCFs in each of the seven ancestral genome reconstructions is larger than 30 (Table 1), whereas the estimated number of chromosomes of those ancestors is 23 or 24 (5, 6, 8, 20, 28, 48). Chromosome fragmentation is caused primarily by large repetitive regions around centromeres and other regions of chromosomes that are difficult to bridge in assemblies that do not have an underlying genetic or physical map. The final step of our reconstruction was to reorder whole reconstructed APCFs of each ancestor on the basis of their predicted ancestral configuration from FISH data (5, 6, 8, 20, 28, 48). To accomplish this, we collected known karyotypes of ancestral genomes predicted by FISH experiments from the literature (5, 6, 8, 20, 28, 48).

Identification of evolutionary breakpoints and chromosome rearrangements

Analysis at 300 Kbp and 500 Kbp resolutions can identify breakpoints caused by translocations, inversions, fissions, fusions, deletions, insertions and transpositions involving SFs of size above these thresholds. Apparent rearrangements involving SFs at higher resolution are possible with DESCHRAMBLER, but at resolutions less than 300 Kbp, presence or absence of breakage in synteny can be affected by assembly errors, alignment artifacts, segmental duplications and copy number variants, leading to an overestimation of the number of chromosome rearrangements. Thus, these algorithmic thresholds yield a conservative definition of evolutionary breakpoints

that capture most of the true chromosomal rearrangements that have occurred during evolution (see below).

Reconstructed ancestral genomes obtained using DESCHRAMBLER are more fragmented than what has been known in part because of scaffold assemblies of descendent species where the exact tips of chromosomes are not known, and in part because of ambiguous cases resulting from insufficient evidence of adjacency. Therefore, ancestral predicted chromosome fragments (APCFs) by DESCHRAMBLER were first reorganized by referring to FISH-based reconstruction results (5, 6, 8, 20, 28, 48), which show large-scale organization of ancestral chromosomes. Then the reorganized APCFs of parent and child ancestors on each branch in a phylogenetic tree were compared to infer the history of the changes of APCFs from the parent to the child ancestor. This process was repeated for branches from the eutherian ancestor to human, and different types of chromosome rearrangements, such as fissions, fusions, inversions, and complex rearrangements (i.e., a combination of inversions and putative transpositions) were identified.

The reconstructed chromosomes of each ancestor were visualized using the Evolution Highway browser (http://eh-demo.ncsa.illinois.edu/ancestors/).

Comparison of chromosome rearrangement rates

Rates of chromosome rearrangement (EBRs/My) were calculated using the number of EBRs detected for each phylogenetic branch divided by the estimated length of each branch (in My) of the tree (4). Only the ancestor rates and the rates on the branches leading to humans and other primates were included in the analysis. The primate lineage was chosen for comparison of rearrangement rates because there is a very high quality reference sequence (human) and it has the greatest number of represented species with chromosome-scale genome assemblies. We estimated rates of chromosome rearrangement at 300 Kbp and 500 Kbp resolution of HSBs. The t-statistics for each branch were obtained by calculating the difference between the rearrangement rate on the branch and the mean rate across all the branches and then normalizing for the standard error. *P*-values were corrected by false discovery rate (FDR) using the p.adjust function from the R package (https://www.R-project.org).

Comparison with existing tools

The reconstructed ancestors of DESCHRAMBLER were compared with results from three existing tools, ANGES (18), inferCARs (3), and MGRA (17). For fair comparison, the four tools were used to predict the adjacencies of the same set of SFs for ancestors, and the similarities and differences of their predicted adjacencies were measured by using the Jaccard index, which is calculated by the number of common adjacencies divided by the union of adjacencies between two sets of adjacencies predicted by two different tools. inferCars was run with default parameters, and MGRA was run with 3 as the number of stages value along with other default parameters. The parameters used for ANGES are shown in Supplementary Table S8. For fair comparison the original reconstruction results obtained using DESCHRAMBLER, not the modified results based on the FISH data, were used.

Evaluation using FISH data

Interchromosomal rearrangements of human chromosomes referenced to computationally reconstructed ancestor chromosomes were identified and compared with reconstructions made using chromosome painting. The FISH-based reconstructions for the eutherian (8), boreoeutherian (6) and simian (20) ancestors were compiled from the literature.

Availability of software and datasets

The source code of DESCHRAMBLER and link to input and output files are available at https://github.com/jkimlab/DESCHRAMBLER.

Acknowledgments

Supported by the Ministry of Science, ICT & Future Planning of Korea Grant 2014M3C9A3063544 (to J.K.); the Ministry of Education of Korea Grant 2016R1D1A1B03930209 (to J.K.); the Rural Development Administration of Korea Grant PJ01040605; the Biotechnology and Biological Sciences Research Council grants BB/K008226/1 and BB/J010170/1 (to D.M.L); the Robert and

Rosabel Osborne Endowment (to H.A.L.); the National Institutes of Health Grant HG007352 (to J.M.) and the National Science Foundation Grants 1054309 and 1262575 (to J.M.)

References

1. White MJD (1969) Chromosomal Rearrangements and Speciation in Animals. *Annu Rev Genet* 3(1):75-98.
2. Murphy WJ*, et al.* (2005) A rhesus macaque radiation hybrid map and comparative analysis with the human genome. *Genomics* 86(4):383-395.
3. Ma J*, et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res* 16(12):1557-1565.
4. Murphy WJ*, et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309(5734):613-617.
5. Ferguson-Smith MA & Trifonov V (2007) Mammalian karyotype evolution. *Nat Rev Genet* 8(12):950-962.
6. Froenicke L (2005) Origins of primate chromosomes - as delineated by Zoo-FISH and alignments of human and mouse draft genome sequences. *Cytogenet Genome Res* 108(1-3):122-138.
7. Fronicke L, Wienberg J, Stone G, Adams L, & Stanyon R (2003) Towards the delineation of the ancestral eutherian genome organization: comparative genome maps of human and the African elephant (Loxodonta africana) generated by chromosome painting. *Proc Biol Sci* 270(1522):1331-1340.
8. Ruiz-Herrera A, Farre M, & Robinson TJ (2012) Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity (Edinb)* 108(1):28-36.
9. Hedges SB, Marin J, Suleski M, Paymer M, & Kumar S (2015) Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* 32(4):835-845.
10. Bovine Genome S*, et al.* (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324(5926):522-528.
11. Larkin DM*, et al.* (2009) Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* 19(5):770-777.
12. Lewin HA, Larkin DM, Pontius J, & O'Brien SJ (2009) Every genome sequence needs a good map. *Genome Res* 19(11):1925-1928.
13. Groenen MA*, et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491(7424):393-398.
14. Bourque G & Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res* 12(1):26-36.
15. Murphy WJ, Bourque G, Tesler G, Pevzner P, & O'Brien SJ (2003) Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum Genomics* 1(1):30-40.
16. Berthelot C, Muffato M, Abecassis J, & Roest Crollius H (2015) The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep* 10(11):1913-1924.
17. Alekseyev MA & Pevzner PA (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res* 19(5):943-957.

18. Jones BR, Rajaraman A, Tannier E, & Chauve C (2012) ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics* 28(18):2388-2390.
19. Kim J, *et al.* (2013) Reference-assisted chromosome assembly. *Proc Natl Acad Sci USA* 110(5):1785-1790.
20. Stanyon R, *et al.* (2008) Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res* 16(1):17-39.
21. Kolmogorov M, Raney B, Paten B, & Pham S (2014) Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 30(12):i302-309.
22. Huddleston J, *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* 24(4):688-696.
23. Mostovoy Y, *et al.* (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* 13(7):587-590.
24. Putnam NH, *et al.* (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 26(3):342-350.
25. Seo JS, *et al.* (2016) De novo assembly and phasing of a Korean human genome. *Nature* 538(7624):243-247.
26. Damas J, *et al.* (2016) Upgrading short read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res* In-press.
27. Froenicke L, *et al.* (2006) Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res* 16(3):306-310.
28. Wienberg J (2004) The evolution of eutherian chromosomes. *Curr Opin Genet Dev* 14(6):657-666.
29. Locke DP, *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529-533.
30. O'Leary MA, *et al.* (2013) The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* 339(6120):662-667.
31. Zhang X, Goodsell J, & Norgren RB, Jr. (2012) Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* 13:206.
32. Zimin AV, *et al.* (2014) A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* 9(1):20.
33. Roberto R, Misceo D, D'Addabbo P, Archidiacono N, & Rocchi M (2008) Refinement of macaque synteny arrangement with respect to the official rheMac2 macaque sequence assembly. *Chromosome Res* 16(7):977-985.
34. Farre M, *et al.* (2016) Novel Insights into Chromosome Evolution in Birds, Archosaurs, and Reptiles. *Genome Biol Evol* 8(8):2442-2451.
35. Rudd MK, *et al.* (2009) Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. *Genome Res* 19(1):33-41.
36. Emerson RO & Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet* 5(1):e1000325.
37. Hayden S, *et al.* (2010) Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res* 20(1):1-9.
38. Ou Z, *et al.* (2011) Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* 21(1):33-46.

39. Duffy DJ (2011) Modulation of Wnt signaling: A route to speciation? *Commun Integr Biol* 4(1):59-61.
40. Harewood L & Fraser P (2014) The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet* 23(R1):R76-82.
41. Mitelman F, Mertens F, & Johansson B (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat Genet* 15 Spec No:417-474.
42. Koepfli KP, Paten B, Genome KCoS, & O'Brien SJ (2015) The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* 3:57-111.
43. Kent WJ*, et al.* (2002) The human genome browser at UCSC. *Genome Res* 12(6):996-1006.
44. Harris R (2007) Improved pairwise alignment of genomic DNA. PhD (Pennsylvania State University).
45. Hedges SB, Dudley J, & Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22(23):2971-2972.
46. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368-376.
47. Sankoff D & Blanchette M (1999) Probability models for genome rearrangements and linear invariants for phylogenetic inference. *Proceedings of the third International Conference on Computational Molecular Biology (RECOMB99)*, pp 302-309.
48. Muller S & Wienberg J (2001) "Bar-coding" primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. *Hum Genet* 109(1):85-94.

Figure 1: Phylogenetic tree of descendant species and reconstructed ancestors. The numbers on branches from the eutherian ancestor to human are the numbers of breakpoints in RACFs, with breakpoint rates (the number of breakpoints per 1 My) in parentheses. The unit of time of branch lengths is 1 My.

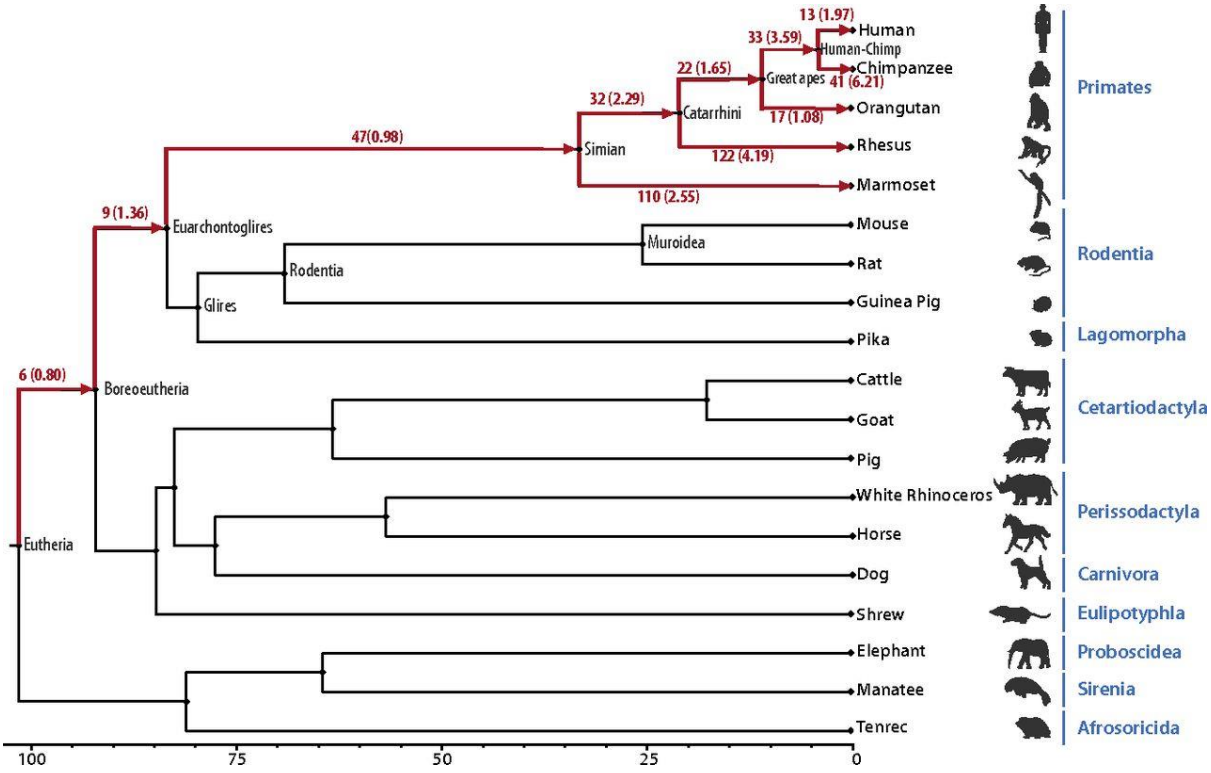

Figure 2: Summary visualization of rearrangements of ancestral eutherian chromosomes in chromosomes of reconstructed descendant ancestors, and extant descendant and outgroup species. Solid red-brown blocks indicate eutherian chromosomes that were maintained as a single synteny block, with shades of the color indicating the fraction of the chromosome affected by intrachromosomal rearrangements (lightest shade is most affected). Split blocks demarcate eutherian chromosomes that were also affected by interchromosomal rearrangements: that is, fissions and translocations. Shades of green in split blocks indicate the fraction of an ancestral chromosome affected by translocations or fissions (lightest shade is most affected), and the shades of red-brown indicate the fraction of eutherian chromosomes affected by intrachromosomal rearrangements measured and summed for all SFs. The heatmap shows the color shades used to represent different fractions of outgroup, descendant ancestors' and extant species chromosomes affected by interchromosomal (shades of green) or intrachromosomal (shades of brown-red) rearrangements. Because of undefined positions of ancestral centromeres, the intrachromosomal rearrangements are measured relative to the prevailing orientation of SFs within each outgroup or descendant chromosome and therefore the fraction of intrachromosomal rearrangements cannot exceed 50%. As it follows from the heatmap, dark shades indicate high level of conservation with the ancestral chromosome and light shades of the same color indicate high level of rearrangements. BOR, boreoeutherian ancestor; CAT, catarrhini ancestor; EUA, euarchontoglires ancestor; EUT, eutherian ancestor; GAP, great apes ancestor; HUC, human–chimp ancestor; SIM, simian ancestor.

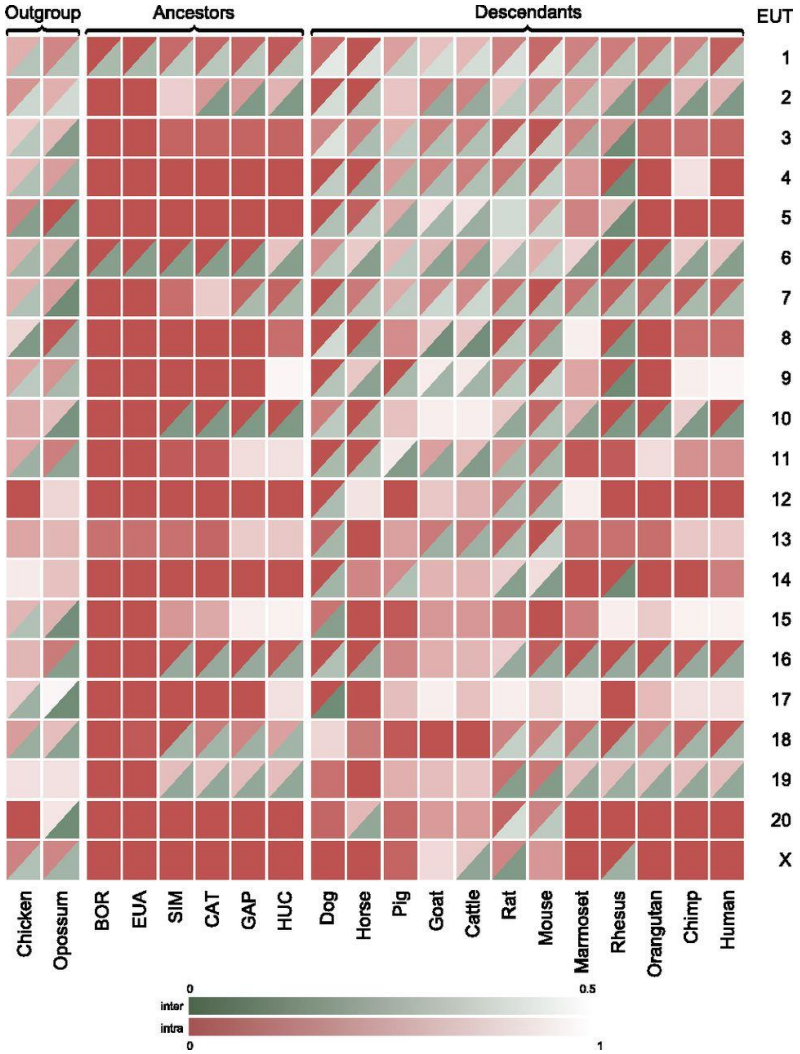

Figure 3: Two examples of eutherian ancestor chromosomes with dramatically different evolutionary histories in the primate lineage. Order and orientation of SFs overlaid on the reconstructed eutherian ancestor chromosomes are visualized using the Evolution Highway comparative chromosome browser (eh-demo.ncsa.illinois.edu/ancestors/). The eutherian chromosome number and its total length are given at the top of each ideogram. Only the main fragment of EUT15 (EUT15a) is shown for this comparison. Blue and pink colors represent orientation of blocks relative to the reference, with blue indicating the same orientation, and pink indicating the opposite orientation. Pink does not always indicate an inversion because the orientation of RACFs is randomly chosen during the reconstruction. Also, as in the case of dog for EUT14, numbering of nucleotides may begin from the opposite end of the chromosome. The number within each block represents a chromosome of a reconstructed ancestor (Dataset S1) or an extant species; a letter indicates a fragment of the chromosome. Adjacency scores computed with DESCHRAMBLER are shown in the right-most tracks. Letter codes of reconstructed ancestors are the same as given in the legend of Fig. 2. Only extant species with full chromosome-scale assemblies are shown. BOR, boreoeutherian ancestor; CAT, catarrhini ancestor; EUA, euarchontoglires ancestor; GAP, great apes ancestor; HUC, human–chimp ancestor; SIM, simian ancestor.